prediction. In order to improve the accuracy of air quality prediction, this study proposes a method based on Elman neural network. This method uses Elman neural network to optimize the prediction results of two air quality numerical models of CMAQ and CAMx. First, this study runs the air quality mode CMAQ and CAMx to get the prediction results, and then pre-process the prediction results. The processed prediction data and the measured data are used as the input of the Elman neural network for model training and finally get the neural network model. Through the verification and analysis of the test data set, the experimental results show that the method shows higher accuracy than the single air quality numerical model.

Key words: Elman neural network; air quality model; result optimization; forecasting of air quality; pollutant concentration

改革开放以来, 我国大力发展工业、制造业, 经济发展的同时, 带来的环境问题也不容忽视. 近年来, 空气质量状况越发得到人们的密切关注, 2017 年我国空气污染状况以华北地区为中心呈放射状分布, 空气质量直接影响到人类的日常生活[1]. 绿水青山就是金山银山, 地处东北老工业基地的沈阳市也面临着同样严峻的空气质量问题. 空气质量的精准预测能够为空气质量的治理提供科技支撑, 各项污染物浓度数据是计算空气质量指数进而衡量空气质量的重要依据.

空气质量数值模式是一种通过大气物理化学方式来模拟污染物之间的相互反应、传输和转化过程, 进而预测空气质量的方法. 空气质量模式 CMAQ 和 CAMx 是依据大气物理化学方法来模拟污染物的扩散和反应过程进而来预测空气质量的方法. 从最初的采用简单线性机制的第一代空气质量模式, 发展到考虑了物质之间的互相作用和相互转化的第三代空气质量模式, 预测精度不断提高. 但是由于空气质量模式受污染源清单数据, 气象数据, 光解文件等输入文件的影响, 输入文件的质量会影响预测结果误差大小. 因此本文提出了一种集成 CMAQ 和 CAMx 两种单一空气质量模式结果的方法, 在单一数值模式的基础上降低误差, 提高预测准确率. 司志娟等[2]将灰色 GM(1, 1) 模型与人工神经网络模型组合, 对天津市 2009 到 2010 的 PM_{10}、SO_2、NO_2 进行预测, 预测相对误差在 5% 以下. 张恒德等[3]利用 BP 神经网络集成了 CUACE、BREMPS 和 WRF-Chem 等 3 个环境模式预报产品, 2015 到 2016 年在北京和石家庄地区污染物浓度和实测值的均方根误差比各单一模式降低了 15% 以上. 梅贵琴[4]利用 Elman 神经网络根据以往臭氧浓度数据预测未来臭氧浓度值, 绝大多数的数据可以达到小于 0.2 的相对误差. 神经网络具有预测未来非线性数据的能力, 应用于各个地区空气质量预测方面取得了良好的效果. 考虑到过往短时间段内空气质量会对未来空气质量产生影响, 而 Elman 神经网络能够增加对过往数据的敏感性. 因此本文提出将 Elman 神经网络用于集成 CMAQ 和 CAMx 两种数值模式的预测结果, 在单一数值模式基础上提高空气质量预测结果的准确度.

1 Elman 神经网络

Elman 神经网络是一种反馈型神经网络, 由输入层, 隐含层, 承接层, 输出层四层组成, 承接层是从隐含层获得反馈信息, 然后再输入到隐含层, 以此来记忆隐含层神经元的上一时刻的输出, 这样的网络结构可以增强对过往数据的敏感度[5–8]. Elman 神经网络结构如图 1 所示. 其中, 输入向量是 r 维的 x 向量, $x=[x_1,x_2,\cdots,x_r]$; 隐含层输出向量是 n 维的 u 向量, $u=[u_1,u_2,\cdots,u_n]$; 输出向量是 m 维的 y 向量, $y=[y_1,y_2,\cdots,y_m]$; 承接层输出向量是 n 维的 x_c 向量, $x_c=[x_{c1},x_{c2},\cdots,x_{cn}]$. $w_{(i,k)},w_{(k,j)},w_{(s,k)}$分别是输入层到隐含层, 隐含层到输出层, 承接层到隐含层的权重矩阵[9]. $f(\cdot)$, $g(\cdot)$分别是隐含层和输出层的激活函数, $h(\cdot)$是承接层激活函数, X_c是承接层输出, t 是时间步长, 输出层输出为:

$$y(t)=g\left(f(t)w_{(k,j)}\right) \tag{1}$$

隐含层输出为:

$$u(t)=f\left(x(t)w_{(i,k)}+x_c(t-1)w_{(s,k)}\right) \tag{2}$$

承接层输出为:

$$x_c = h(u(t-1)) \tag{3}$$

Elman 神经网络模型的算法学习流程, 首先要初始化各层节点的权值, 然后输入训练数据, 计算各层的输入输出值. 其中将隐含层上一轮的输出, 输入到承接层, 数据经过承接层处理后在本轮和输入层数据一同输入到隐含层. 最后根据输出层的结果和误差函数计算误差, 若误差的大小满足要求或训练次数达到最大, 则停止训练, 否则更新权值, 进入下一轮训练.

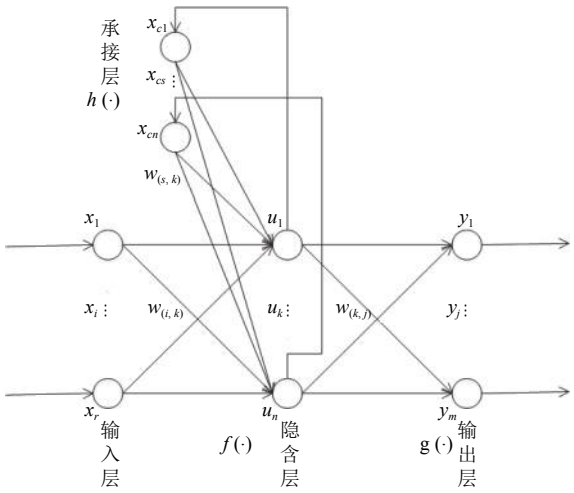

图 1 Elman 神经网络结构示意图

误差函数为:

$$E(t) = \frac{1}{2}(y(t) - y_a(t))^{\mathrm{T}}(y(t) - y_a(t)) \tag{4}$$

式 (4) 中, $y_a(t)$是标准实际输出数据, $y(t)$模型输出数据.

根据误差逆向传播算法, 得到:

$$\Delta w_{(k,j)} = \eta_3 \delta_j u_k(t) \tag{5}$$

$$\Delta w_{(i,k)} = \eta_2 \delta_k x_{cs}(t) \tag{6}$$

$$\Delta w_{(s,k)} = \eta_1 \sum_{k=1}^{n} \left(\delta_j w_{(k,j)}\right) \frac{\partial x_{cs}(t)}{\partial w_{(s,k)}} \tag{7}$$

其中,

$$\delta_j = \left(y_j(t) - y_{aj}(t)\right) g'_j(.) \tag{8}$$

$$\delta_k = \sum_{j=1}^{m} \left(\delta_j w_{(k,j)}\right) f'_k(.) \tag{9}$$

$$\frac{\partial x_{cs}(t)}{\partial w_{(s,k)}} = f'_k(.)\, x_{cs}(t-1) + \alpha \frac{\partial x_{cs}(t-1)}{\partial w_{(s,k)}} \tag{10}$$

η_1, η_2, η_3分别是$w_{(i,k)}$, $w_{(k,j)}$, $w_{(s,k)}$的学习率, δ_j是输出层神经元的梯度项, δ_k是隐含层神经元的梯度项[10]. x_{cs}是承接层第s维输出, u_k是隐含层第k维输出. $y_j(t)$是第j个结点第t轮的输出值, $y_{aj}(t)$是第t轮第j个结点的标准输出值. $g'_j(.)$是输出层的导数, $f'_k(.)$是隐含层的导数, $0 \le \alpha < 1$.

学习算法伪代码如算法 1 所示.

算法 1. 学习算法. 训练集$D=\{(x_t, y_t)\}$

```
1. 初始化 Elman 神经网络中所有连接权重和阈值.
2. for all (x_t,y_t)∈D do
3. 根据式 (1)~式 (3) 计算每一层的输出值;
4.     根据式 (8) 计算输出层神经元梯度项;
5.     根据式 (9) 计算隐含层神经元梯度项;
6.     根据式 (5)~式 (7) 更新权值;
7.     if(达到停止条件)
8.         break;
9.     end if
10. end for
```

2 实验分析

本文数据集来源为空气质量数值模式 CMAQ 和 CAMx 输出的辽宁省沈阳市 6 项常规污染物 (包括 $PM_{2.5}$、PM_{10}、SO_2、CO、NO_2、O_3) 的浓度结果, 以及在中国空气质量在线监测分析平台所下载的 6 项常规污染物的实测数据. 本文中的数据集的大小为 2019 年 1 月到 2019 年 6 月共 181 天, 起报时刻为 20 时, 预报时长为未来 4 天的 6 项常规污染物 24 小时平均浓度数据, 其中选取 30 条数据用作测试数据来评价模型的效果, 剩余数据用于训练模型.

2.1 数据预处理

空气质量模式受气象数据, 地理数据, 以及污染源清单数据等输入文件的影响, 会出现数据缺测情况, 首先要去除缺测值, 减少缺测数据对模型训练的影响.

为了减少不同量纲对后续数据分析和模型训练造成影响, 需要先采用 Min-Max 线性归一化方法对数据进行归一化处理, 将原数据映射到 (0, 1) 之间, 公式如式 (11) 所示:

$$x_{\mathrm{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{11}$$

其中, x_{norm}表示归一化后的数据, x表示原数据, $x_{\min}$表示的是数据集中的最小值, $x_{\max}$表示的是数据集中的最大值.

2.2 实验过程

实验过程如图 2 所示, 首先运行空气质量模式 CMAQ 和 CAMx, 对得到的空气质量模式预测结果去除缺测数据处理, 然后将空气质量模式预测结果和实测数据进行归一化处理, 处理后的数据作为 Elman 神经网络的输入, 初始化各层结点的权值, 进行训练.

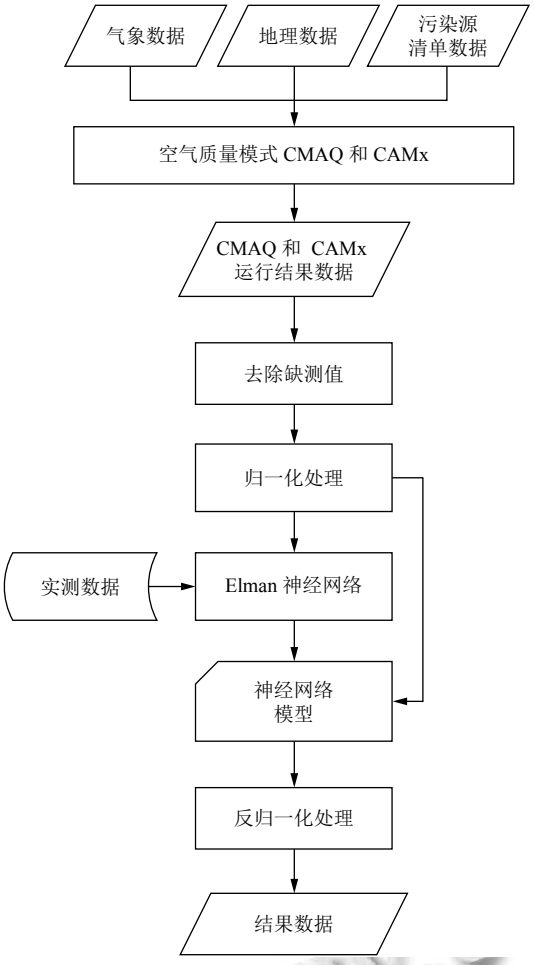

图 2 实验流程图

由于本算法需要结合 CMAQ 和 CAMx 两个模型优化预测值, 所以把输入层节点数设置为 2, 输出层节点数设置为 1, 经过敏感性实验得到隐含层节点数设置为 10 时效果最佳; 输出层激活函数设置为输入和输出相等的 Purelin 函数; 隐含层激活函数设置为 Sigmoid 函数, 其公式如下所示:

$$f(x)=\frac{1}{1+e^{-x}} \tag{12}$$

当训练误差达到最小 (0.01) 或达到最大训练轮数 (10 000) 停止训练. 将测试数据输入到训练好的模型中, 得到模型输出结果, 对模型输出结果进行反归一化处理, 得到 Elman 神经网络模型优化的 CMAQ 和 CAMx 预测结果.

2.3 评价指标

本文使用均方根误差 (*RMSE*), 平均绝对误差 (*MAD*), 和平均绝对百分比误差 (*MAPE*), 来定量分析模型结果的精度[11–14]. 3 个评价指标公式如下:

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{ai}-y_i)^2} \tag{13}$$

$$MAD=\frac{1}{N}\sum_{i=1}^{N}|(y_{ai}-y_i)| \tag{14}$$

$$MAPE=\frac{\sum_{i=1}^{N}\frac{|y_{ai}-y_i|}{y_{ai}}}{N}\times 100\% \tag{15}$$

其中, y_i是预测值, y_{ai}是实测值. 均方根误差 (*RMSE*) 是预测值与实测值误差的平方和与实验次数 N 比值的平方根, 它能反应出预测值和实测值的偏差大小以及预测结果的稳定程度. 平均绝对误差 (*MAD*) 是预测值和实测值的绝对误差和与实验次数 N 的比值. 平均绝对百分比误差 (*MAPE*) 能反应模型的优劣程度, 是相对误差的和与试验次数的比值.

2.4 结果分析

图 3 是对比实验结果图, 横坐标表示时间序列, 纵坐标表示污染物浓度, 从图 3 中可以看到, CMAQ 和 CAMx 两个单一数值模式的 SO_2 和 $PM_{2.5}$ 预测结果偏高, NO_2 预测峰值时结果偏高, PM_{10} 和 CO 预测趋势和实测值大致相同, 个别地方相差较大; 而经过 Elman 神经网络优化的预测结果与单一模式相比较更为接近实际值.

空气质量模式 CMAQ, CAMx 和基于 Elman 神经网络集成后结果的评价指标对比情况, 如表 1 所示. 综合表 1 和图 3 我们可以看到, 就沈阳市的预报结果而言, 两种单一模式对于 6 项污染物的预测结果都有不同程度的误差. 而集成后的 6 项污染物预测结果的预测误差有所下降, 综合对比 3 个评价指标可以看到 $PM_{2.5}$ 和 SO_2 的 *MAD* 和 *RMSE* 都有下降, *MAPE* 有大

幅度下降, 其中 $PM_{2.5}$ 的 *MAPE* 下降了 50%–88%, SO_2 的 *MAPE* 下降了 110%–209%; 而 PM_{10}, CO, NO_2 和 O_3 的 *MAD* 和 *RMSE* 有小幅度下降或持平, *MAPE* 都有所下降. 总体来说, 基于 Elman 神经网络优化两种单一空气质量模式的结果相比于单一空气质量模式的预测精度和稳定性有所提高.

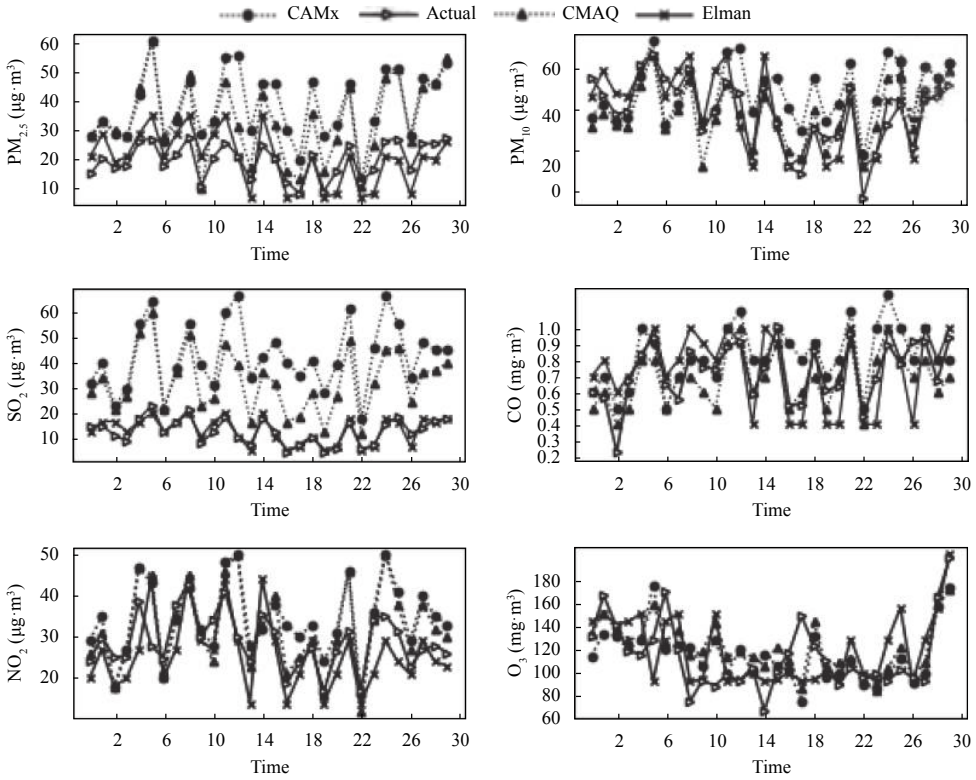

图 3 实验结果图

表 1 实验结果对比

评价指标	$PM_{2.5}$ (μg/m^3)			PM_{10} (μg/m^3)			SO_2 (μg/m^3)		
	CMAQ	CAMx	集成	CMAQ	CAMx	集成	CMAQ	CAMx	集成
RMSE	16	20	10	13	18	12	21	31	7
MAD	14	18	9	11	16	11	19	29	6
MAPE (%)	86	124	36	32	62	31	152	254	42
评价指标	CO (mg/m^3)			NO_2 (μg/m^3)			O_3 (μg/m^3)		
	CMAQ	CAMx	集成	CMAQ	CAMx	集成	CMAQ	CAMx	集成
RMSE	0.207	0.243	0.201	10	11	8	23	26	22
MAD	0.177	0.190	0.173	8	8	7	18	20	18
MAPE (%)	29	36	23	31	41	27	17	17	14

3 结论

本文针对空气质量预测提出了在 CMAQ 和 CAMx 两个空气质量数值模型基础上, 通过 Elman 神经网络集成两个数值模式结果的方法. 实验结果表明, 本文提出的方法结合了两种模型的优势, 提高了预测精度和稳定性, 降低了单一空气质量数值模式的预测误差, 从而能够为后续空气质量预报以及空气质量控制提供数值依据.

参考文献

1 谷金科. 中国城市空气质量影响因素研究[硕士学位论文]. 太原: 山西财经大学, 2019.

2 司志娟. 基于灰色神经网络组合模型的空气质量预测[硕士学位论文]. 天津: 天津大学, 2012.
3 张恒德, 张庭玉, 李涛, 等. 基于 BP 神经网络的污染物浓度多模式集成预报. 中国环境科学, 2018, 38(4): 1243–1256.
4 梅贵琴. 改进的 Elman 神经网络和网络参数优化算法研究[硕士学位论文]. 重庆: 西南大学, 2017.
5 李丹. 基于神经网络的北京联通 GSM 话务量预测[硕士学位论文]. 北京: 北京邮电大学, 2009.
6 林春燕, 朱东华. 基于 Elman 神经网络的股票价格预测研究. 计算机应用, 2006, 26(2): 476–477, 484.
7 韩旭明. Elman 神经网络的应用研究[硕士学位论文]. 天津: 天津大学, 2006.
8 Zhang Y, Wang XP, Tang HM. An improved elman neural network with piecewise weighted gradient for time series prediction. Neurocomputing, 2019, 359: 199–208. [doi: 10.1016/j.neucom.2019.06.001]
9 夏杨. 基于改进型 Elman 神经网络的电力负荷预测[硕士学位论文]. 西安: 西安理工大学, 2017.
10 周志华. 机器学习. 北京: 清华大学出版社, 2016.
11 江琪, 王飞, 张恒德, 等. 北京市 $PM_{2.5}$ 和反应性气体浓度的变化特征及其与气象条件的关系. 中国环境科学, 2017, 37(3): 829–837.
12 闫以聪. 回归方程与神经网络在数值预测方面的对比研究综述. 数理医药学杂志, 2007, 20(1): 66–69.
13 姚文强. 基于数值预报的空气质量预测模型的研究[硕士学位论文]. 杭州: 浙江理工大学, 2017.
14 Rotich NK, Backman J, Linnanen L, *et al*. Wind resource assessment and forecast planning with neural networks. Journal of Sustainable Development of Energy, Water and Environment Systems, 2014, 2(2): 174–190. [doi: 10.13044/j.sdewes.2014.02.0015]